DOCUMENT RESUME

ED 456 140 TM 033 202

AUTHOR	Horn, Catherine; Ramos, Miguel; Blumer, Irwin; Madaus, George
TITLE	Cut Scores: Results May Vary. NBETPP Monographs, Volume 1, Number 1.
INSTITUTION	National Board on Educational Testing and Public Policy, Chestnut Hill, MA.
SPONS AGENCY	Ford Foundation, New York, NY.
PUB DATE	2000-04-00
NOTE	33p.
AVAILABLE FROM	For full text: http://www.nbetpp.bc.edu.
PUB TYPE	Opinion Papers (120) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Standards; *Cutting Scores; *Decision Making; Elementary Secondary Education
IDENTIFIERS	Angoff Methods; Massachusetts Comprehensive Assessment System; *Standard Setting

ABSTRACT

This paper discusses how cut scores are set and used and how accurately they reflect student achievement. Regardless of the method used, the cut-score setting process is subjective. The cut score is the point on a score scale that separates one performance standard from another. Cut scores may also be used to set performance levels for open-response assessments like essay tests. This monograph discusses three methods of setting cut scores: (1) the modified Angoff method; (2) Contrasting Groups; and (3) Bookmark procedures. A description of each process shows the pros and cons of each approach, especially when the scores are used to make high stakes decisions. All of these methods are subjective to some degree; it is important not to assume that established cut scores accurately reflect student performance. Other external evidence can help establish whether the choices made are appropriate. An example is provided in which students' performance on a second commercially developed standardized test is used to examine whether or not the performance standards of one high-stakes state examination, the Massachusetts Comprehensive Assessment System (MCAS), are capricious. The MCAS appears to tap into information that is similar, but not identical, to other standardized tests. Several possible reasons for the unexplained variance are discussed. It is important to remember that the MCAS is not an unquestionable source of information about student performance and that different cut scores might provide different information to students about their accomplishments. In educational assessment, there is the fundamental problem that performance levels are based on cut scores, and cut scores are based on judgment. (Contains 11 endnotes.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 456 140

The National Board on Educational Testing and Public Policy

NBETPP

monographs

Volume 1 Number 1 April 2000

Cut Scores: Results May Vary

Catherine Horn, Miguel Ramos, Irwin Blumer, and George Madaus
National Board on Educational Testing and Public Policy
Peter S. and Carolyn A. Lynch School of Education
Boston College

Increasingly, states are holding students accountable for learning through tests. Often, students must pass tests in order to achieve specific milestones such as graduation from high school or promotion to the next grade. But what does "pass" mean? In simple terms, passing a test is scoring above a predetermined level. A cut score is the point that sets that predetermined level; it differentiates between those who pass and those who fail. Multiple cut scores may also be set to more finely separate students into categories. For example, students may be classified as advanced, proficient, needs improvement, or unsatisfactory. This paper addresses two issues. How are cut scores set and used? And how accurately do they reflect student achievement? Regardless of the method, the cut-score setting process is subjective. It is important to understand how cut scores are established in order to evaluate how they are used.

TM033202

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
M. Clarke
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

What is a Performance Standard?

A performance standard or performance level describes a specific level of competence. Some common categories include Proficient, Advanced, and Unsatisfactory. Exactly what constitutes "proficient," "unsatisfactory" or other categories depends largely on policy makers' decisions. Figure 1 shows the performance standards (referred to here as achievement ratings) for a state test in Louisiana (LEAP 21). Each performance standard is briefly characterized. But how do we know what separates Basic achievement from Proficient? This is where the cut score comes into play.

Figure 1

Performance Level Definitions

LEAP 21 Achievement Ratings

Advanced: Superior performance

Proficient: Competency over challenging subject matter

Basic: Mastery of only fundamental knowledge and skills

Approaching Basic: Partially demonstrates fundamental knowledge and skills

Unsatisfactory: Does not demonstrate fundamental knowledge and skills

The cut score is the point on a score scale that separates one performance standard from another. Theoretically, cut scores can be used to create any number of divisions. The most common categorizations range from pass/fail to a five-standard set like the Louisiana example. Figure 2 shows five performance levels and corresponding test scores on a 50-question multiple-choice (MC) test. As we can see, each achievement rating has a specific range of correct responses. Thus, a student with between 30-39 correct answers would be categorized as Proficient while a student with between 20-29 correct would be labeled Basic.

Cut scores may also be used to set performance levels for open-response assessments like essay tests. Ultimately, the purpose of the cut score remains that same. That is, it separates one group of students from another based on predetermined ideas of student achievement. We next discuss a few of the methods policy makers can use to arrive at these distinctions. It is important to remember that each of these methods relies on some form of judgment.

Figure 2

Sample of Cut Scores and Corresponding Percentage of Items Correct

Performance Level	Number of Items Correct	Percent Correct
Advanced	40-50	80-100
Proficient	30-39	60-79
Basic	20-29	40-59
Approaching Basic	10-19	20-39
Unsatisfactory	0-9	0-19

Standard Setting Procedures

Of the several methods available for setting cut scores, we will concentrate on three. These are the Modified Angoff, Contrasting Groups, and Bookmark procedures. A description of the process used in each will help to illustrate how policy makers go about the business of setting cut scores for various assessment programs. It will also allow us to weigh the pros and cons of such procedures, especially when the scores are used to make high stakes decisions, for example, about graduation from high school or promotion.

When we talk about standard setting, people may get the impression that the process is objective. This is not the case. Instead, procedures currently used to establish cut scores rely on some sort of judgment. Thus the results obtained from a well-conceived standard setting procedure may still be suspect if the judgments are made by individuals with inadequate qualifications. Many states try to address this concern by selecting judges from different fields that have some bearing on the process as a whole. The composition of the Virginia Standard Setting Committees illustrates this point. The committees included teachers, curriculum experts, and educators from throughout Virginia and reflected a balance of geographic distribution, ethnicity and race, and knowledge of the grades and content areas to be tested. Each committee had approximately 20 members, nominated by school division superintendents, educational organizations, institutions of higher learning, and the

Catherine Horn is a doctoral student in the Educational Research, Measurement, and Evaluation program in the Lynch School of Education at Boston College.

Miguel Ramos is a doctoral student in the Educational Research, Measurement, and Evaluation program in the Lynch School of Education at Boston College.

Irwin Blumer is a Senior Fellow with the National Board on Educational Testing and Public Policy and a Professor in the Lynch School of Education at Boston College.

George Madaus is a Senior Fellow with the National Board on Educational Testing and Public Policy and the Boisi Professor of Education and Public Policy in the Lynch School of Education at Boston College.

business community. Superintendents representing each of the eight Superintendent Regions in the state chaired the committees. Of course, a varied membership on a standard setting committee does not insure appropriateness of the standards that result. Nothing can absolutely guarantee that. However, the Virginia example provides a credible attempt in that direction. We now turn to the individual cut-score setting procedures.

Angoff and Modified Angoff Methods

The original Angoff procedure requires experts to determine the probability of a minimally competent person answering a particular test question correctly. Assuming the judges are well qualified – with training or background in the domain being tested – you could think of this as a well educated guess. In Figure 3 below, a judge has determined that a minimally competent person has a 95 % chance of getting question 1 right; or, put another way, of a group of 100 minimally competent examinees, 95 would get the question right. Each probability in column 2 represents the expert opinion of the same judge for the

Figure 3

Examples of Calculations for Angoff's Method

Question	Probability of Correct Answer
1	.95
2	.80
3	.90
4	.60
5	.75
6	.40
7	.50
8	.25
9	.25
10	.40
	Sum=5.80

Source: Livingston, S. and Zieky, M. (1982). Methods based on judgements about test questions. In *Passsing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

relevant question. The total at the bottom is the sum of these probabilities. If we were trying to separate minimally competent individuals from incompetent ones, this judge would recommend a cut score at 5.80.

There is one more step. In the interest of fair play, the decision on where to set the cut score is not left to one person. Instead, several experts are asked to weigh in with their judgments. If ten experts were involved, each would be asked to make the probabilistic judgments shown in Figure 3. Thus, we would find ten sums instead of one (Figure 4). The cut score would simply be the average of the sums of all the judges.

Figure 4

Setting the Cut Score Using Angoff's Method

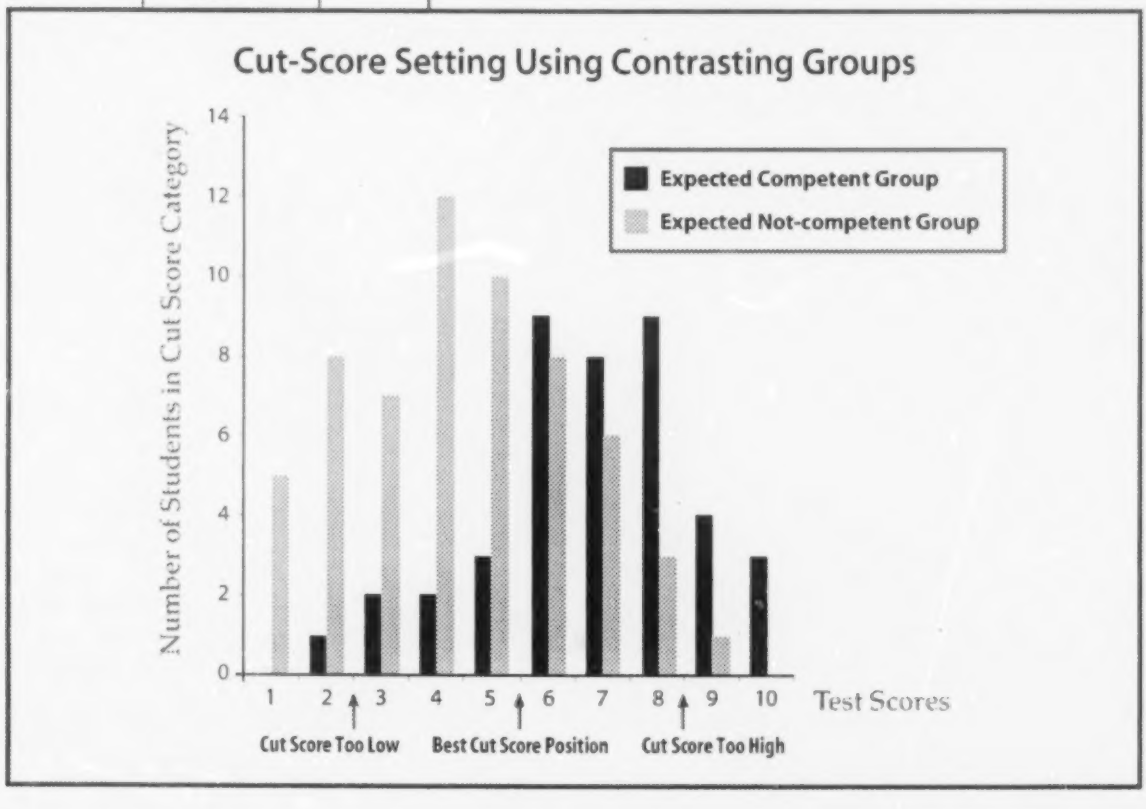
Judge	Cut Score
1	5.80
2	6.00
3	6.00
4	5.40
5	5.00
6	5.30
7	5.50
8	4.80
9	6.10
10	5.50
Average Cut Score for Minimum Competency	**5.54**

The key difference between the original and the Modified Angoff methods lies in the ability of judges to make their own determinations of probability. In the original method, experts determine probability on their own. In theory, a judge could select any probability from 0 to 1.0 (e.g. .95, .43, .55). The Modified Angoff method restricts these probabilities to eight choices (.5, .20, .40, .60, .75, .90, .95) including "Do not know." The cut scores are then determined as in the original Angoff method.

Contrasting Groups

As the name implies, the Contrasting Groups standard setting procedure calls for a comparison of expected performance versus actual performance among different ability groups. The cut score is based on these comparisons. Let us suppose we are trying to determine the cut score for a 10th grade math test with two performance standards (competent and not competent). We first gather a random sample of all tenth grade examinees. Once this group is selected, experts then separate students into two ability levels based on some external factor like course grades or types of courses taken. Students who are enrolled in higher-level courses might be placed in the "expected competent" category and those students in lower-level courses in the "expected not-competent" category. Each student is then labeled accordingly and the test is administered. Test scores for each group are plotted on a graph. We can now compare individuals' expected competency with their actual competency level (as defined by test scores).

Figure 5



From the graph (Figure 5), we see that the "expected competent" students tend to cluster around the higher scores while the "expected not competent" students fall toward the lower scores. Judges now determine what score best discriminates between the "competent" and the "not competent" students. If the cut score is set too high, too many "expected competent" students would be labeled as "not competent." If it is set too low, the opposite is the case. The best cut score is the point that minimizes the number of mis-classified individuals. Note that this graph represents an idealized version of what should occur if the external factor (i.e. courses taken or grades) accurately differentiates between the Contrasting Groups. In fact, we have made several other critical assumptions as well: that judges could distinguish between minimal competency and non-competency; and that the test itself is an accurate measure of 10th grade math achievement. This level of accuracy will rarely be achieved, if ever.

The Bookmark Procedure

The Bookmark procedure was developed by CTB/McGraw-Hill in 1996. Using an arcane statistical technique called item response theory, test questions are ordered along a scale of difficulty, from easy to hard. The judges then determine where cut scores should be set along that scale. Let's say we want to create three performance standards: Advanced, Proficient, and Unsatisfactory. Judges receive a description of what someone at each performance level should be able to accomplish and set cut scores at points they think appropriate– that is, "bookmarking" the items they believe separate one ability level from another. Here again, choices made are based on judgment.

Like the previous methods, the Bookmark procedure does not leave the final decision to one judge. Instead, judges confer about their scores to see if there are any discrepancies. This step helps judges to arrive at a consensus on the performance standard definitions. Judges then repeat the Bookmark procedure. After the second round, they evaluate what they think students should know at each level. A third round gives the judges a chance to amend or defend their final cut score decisions. The average score for each level is then calculated. This represents the final set of cut scores for the exam.

In understanding these procedures, it is important to note that states often employ a combination of cut score procedures rather than just one. Virginia uses components of both the Bookmark and modified Angoff methods. Similarly, Massachusetts uses the Student-based Constructed Response Method, a technique with features characteristic of all three methods discussed previously.

Multiple Performance Levels and Open-Response Questions

While the demand for more performance levels increases, the methods for establishing them remain virtually unchanged.

While the demand for more performance levels increases, the methods for establishing them remain virtually unchanged. Each of the preceding standard setting methods can be used to create any number of performance levels. In the Angoff or Modified Angoff, judges ask themselves what the probability is of an student answering a particular question correctly. They merely need a description of the desired performance levels. In Contrasting Groups too, judges could create three categories as easily as two using the same procedure. The same is true of the Bookmark performance levels. To add a fourth category we would simply describe the new standard and judges would decide what a minimally competent person should be expected to know.

These methods may also be used to create cut scores for open-response items. In each method, the performance standard for a particular item or set of items would need to be established. If we wanted a three point system, we would need three sample essays to represent different levels of competency. The "best" essay would receive the highest rating; the next "best" essay would receive a lower rating and so on until each level was represented. Once these categorizations were established, judges would then determine the cut score based on these ratings.

In the Modified Angoff method, experts might be asked to judge the probability of a minimally competent examinee getting a rating of 2 or 3. As before, the average of several judges' ratings would be used to establish the cut score for the essay. For instance, if a majority of the judges felt that a minimally competent student should be able to achieve at least a 2 on the essay, this would be the cut score for that particular item.

Using our previous example for Contrasting Groups, judges in this procedure would simply have to determine which rating best discriminated between the "competent" and "not competent" groups.

Open-response questions also have a place in the Bookmark procedure. Rather than making a separate judgment for open-response items, the Bookmark procedure incorporates essay ratings into its difficulty scale. Let us use our three-point essay scale. As stated in the second paragraph of this section, each point on our rating scale represents a different competency level. If we assume higher competency levels are more difficult to achieve, we can think in terms of our original difficulty scale. The lowest essay ratings would be found near the bottom while the highest ratings would be near the top. For example, a rating of 1 would be found among the easiest questions. Thus, the essay ratings are incorporated into the overall difficulty scale for the test. That is, a student who is likely to get the easiest questions correct would also be expected to receive an essay rating of at least 1. As the difficulty level increases so does the essay rating. In this manner, determination of the cut score then follows the same procedure detailed in the Bookmark section of this paper.

How Are Performance Levels Used?

The trend toward multilevel performance standards has been accompanied by a wave of reform efforts wrapped in the blanket of accountability – notably the push for more "high-stakes" testing. The term "high stakes" applies not only to students, but also to teachers, schools and districts. For example, beginning with the class of 2004, Virginia high school students will have to pass six of the state Standards of Learning exams in order to graduate. The Washington Post reports that at McClean High School, one of the highest-performing schools in the state, at least 27 percent of the students would have failed to earn their diploma if the graduation requirement had been in effect in 1999. In addition, by 2007, a school in Virginia with less than 70 percent of its students passing the tests in each of four basic subjects will lose its accreditation. But only 6.5 percent of schools in Virginia met the state performance targets on the latest battery of tests; more than 80 percent failed

The trend toward multi-level performance standards has been accompanied by a wave of reform efforts wrapped in the blanket of accountability – notably the push for more "high-stakes" testing.

in at least two subjects. If these results are any indication, there is great cause for concern.

The use of performance levels is criticized chiefly for the arbitrariness of the standard setting procedures. One critic states that, no matter what procedures are used, judgments of performance remain arbitrary: "those who claim the ability to [determine] mastery or competence in statistical or psychological ways...cannot determine 'criterion levels' or standards other than arbitrarily."[1] Because performance levels are based on *judgment*, there is no certainty that the result – the cut scores – reflects any objective measure of competency. Others believe, however, that performance standards can be a worthwhile endeavor if the decisions are made by persons familiar with sound standard setting procedures. They suggest "that, although passing scores are arbitrary in the sense that they are based on judgment, they do not have to be arbitrary in the sense of being capricious."[2]

Cut Scores and Political Expedience

In the case of Richardson v. Lamar County Board of Education, plaintiff Richardson contested the Board's decision not to renew her teaching contract because she failed to pass the Alabama Initial Teacher Certification Tests. In ruling for the plaintiff, Judge Myron Thompson concluded that the method of determining the cut-scores was "so riddled with errors that it can only be characterized as capricious and arbitrary."

> The cut scores...were so astoundingly high that they signaled...an absence of correlation to minimum competence. For example, of the more than 500 teachers who took the first administration of the core examination, none would have passed.
>
> Faced with this problem, the test developer made various mathematical "adjustments" to the original cut score...The State Department of Education was then given the option of lowering the cut scores...which clearly, even after the various adjustments...were not measuring competence...Instead of challenging what the developer had done, the state simply lowered the cut score...to arrive at a "politically" acceptable pass rate.
>
> The State Board of Education and the test developer in effect abandoned their cut-score methodology, with the result that arbitrariness, and not competence, became the touchstone for standing setting.

Source: The National Committee on Testing and Public Policy (1990). *From Gatekeeper to Gateway: Transforming Testing in America*. Chestnut Hill, MA: National Committee on Testing and Public Policy.

BEST COPY AVAILABLE

It is important we not simply accept that the established cut scores accurately reflect student performance. Other external evidence can help to establish whether the choices are appropriate. In the next section, we will use students' performance on a second commercially developed standardized test to examine whether or not the performance standards of one high-stakes state exam, the Massachusetts Comprehensive Assessment System (MCAS), are capricious.

Performance Levels and the MCAS

Massachusetts implemented the MCAS in the spring of 1998. It measures students' performance in the domains of English/Language Arts, Mathematics, and Science/Technology. Each test contains multiple-choice, short-answer, and open-response questions. Where a student's score falls along a range of scale scores from 200 to 280 determines that student's performance level: Failure (200-220), Needs Improvement (221-240), Proficient (241-260), and Advanced (261-280). For example, a student might receive a 236 on the Science/Technology section, receiving the label "needs improvement". The performance levels reduce the 80 potential scale scores to four cut scores.

In the fall of 1998, students, parents, teachers, schools, and the public got their first glimpse at how Massachusetts students fared in reference to high standards for academic achievement as defined by the state in their curriculum frameworks. The numbers were troubling: 81 percent of the fourth graders were failing or in need of improvement in English/Language Arts; 71 percent of the eighth graders fared as poorly in Science/Technology; 74 percent of the tenth graders failed or needed improvement in Mathematics. The message politicians and the public took from the results was clear: the state's education system needed serious reform.

Because of the arbitrary nature of cut-score setting discussed previously, the validity of these cut scores must be considered carefully. That is, how accurate are the inferences and descriptions based on them? One measure of validity is how well performance on one test correlates with that on another test of the same domain. We obtained MCAS scores as well as scores (in the form of national percentile ranks) on other national standardized tests from four districts in Massachusetts.[3] District A provided Stanford 9 scores. The Stanford 9 is a nationally normed test that measures student achievement in, among other things, math, reading, and language arts. District B furnished student scores on the Explore, a test produced by the American College Testing Service (ACT). District C shared their Educational Records Bureau (ERB) exam results. The ERB is a historically difficult test used primarily in private schools and wealthy suburban districts. Finally, District D provided Preliminary Scholastic Aptitude Test (PSAT) scores for students that took the exam in the 10th grade. The four districts vary socioeconomically as well as geographically. They are described in Figure 6.

Figure 6

Characteristics of Four School Districts

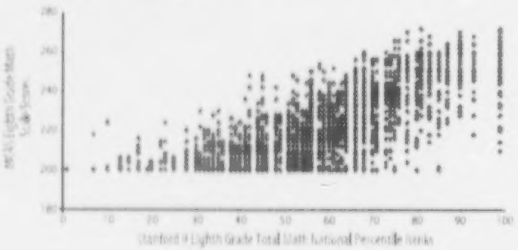
District	Description	Second Standardized Test Used	Grade	Number of Students Tested
A	Low-income and working class urban	Stanford 9 Math	8	3728
B	Working class suburban	Explore Science	8	176
C	Wealthy suburban	ERB Reading	4	153
D	Wealthy urban	PSAT Math	10	287

The graphs for the districts' MCAS scale scores and second standardized tests are presented in Figure 7. As can be seen, each commercially developed standardized test has a positive relationship with the MCAS. A comparison between the MCAS and the Stanford 9 tests for the 3,728 8th grade participants in District A, for example, shows a positive relationship between the two tests. Generally, students with low scores on the MCAS math test have low scores on the Stanford 9 math exam, and, conversely, those with high scores on the MCAS have high scores on the Stanford 9.

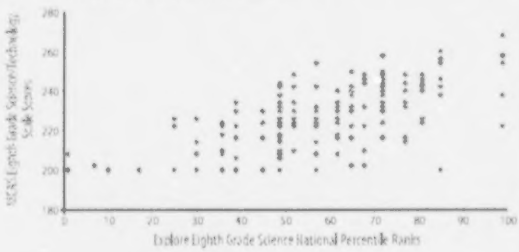
Figure 7

MCAS Scale Scores Versus National Percentile Ranks, by Grade and District

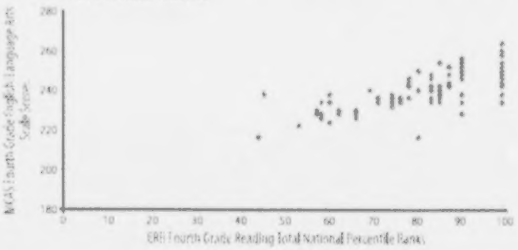
MCAS Eighth Grade Math Scale Scores Against Stanford 9 Eighth Grade Total Math National Percentile Ranks



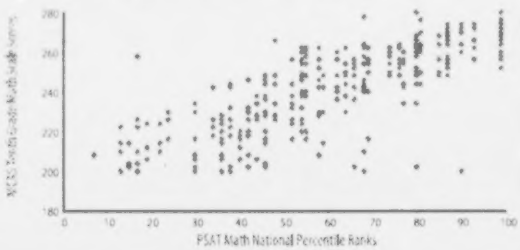
MCAS Eighth Grade Science/Technology Scale Scores Against Explore Eighth Grade Science National Percentile Ranks



MCAS Fourth Grade English / Language Arts Scale Scores Against ERB 4th Grade Total Reading National Percentile Ranks



MCAS Tenth Grade Math Scale Scores Against PSAT Math National Percentile Ranks



BEST COPY AVAILABLE

These positive relationships can also be expressed as correlations (see Figure 8).[4] The correlation between the Stanford 9 and MCAS math scale scores is .84. This correlation indicates that students who do well on the Stanford 9 test also tend to do well on the MCAS math test. The Explore science and the MCAS Science/Technology scores for the eighth graders have a correlation of .67. Although this is lower than ideal, it still indicates similarity in the two test domains.[5] The correlation of .69 between the MCAS reading and the ERB reading scores is again low, but an acceptable indication that the two tests are measuring similar bodies of knowledge. Finally, the correlation between the PSAT and MCAS math scores is .76, suggesting similar domains. The correlations for all these tests indicate, then, that the MCAS reasonably predicts performance on another test.

Figure 8 also presents the correlations between the four MCAS performance levels (i.e., Failure, Needs Improvement, Proficient, Advanced) and the national percentile ranks for the other standardized tests. The correlations for these compari-

Figure 8

Characteristics of Four School Districts

District	Second Standardized Test Used	Correlation between MCAS Scale Scores and Second Test	Variance Accounted for in Scale Scores	Correlation between MCAS Performance Levels and Second Test	Variance Accounted for in Performance Levels
A	Stanford 9 Math	.84	.70	.78	.61
B	Explore	.67	.45	.60	.36
C	ERB Reading	.69	.48	.58	.34
D	PSAT Math	.76	.58	.74	.55

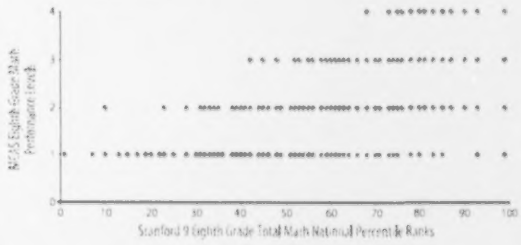
sons remain positive but are lower than those between the MCAS scale score and a second measure. This is due to the variety of scores that can be obtained. Because the possible scores on the MCAS are reduced from 80 points when scale scores are used to 4 points when performance levels are used, the correlation is smaller. Whether we use scale scores or performance levels, however the MCAS tests are not perfectly correlated with the other standardized tests. A student's performance on one test does not exactly match his or her performance on a second test.

Figure 9 presents graphs of the four districts' MCAS performance level results plotted against the national percentile ranks of their second standardized tests. Note that students with

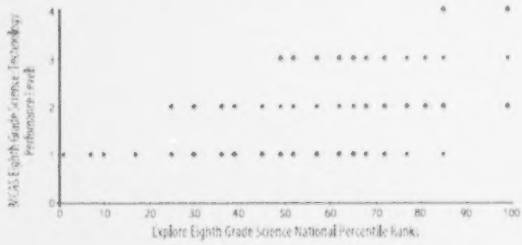
Figure 9

MCAS Performance Levels Versus National Percentile Ranks, by Grade and District

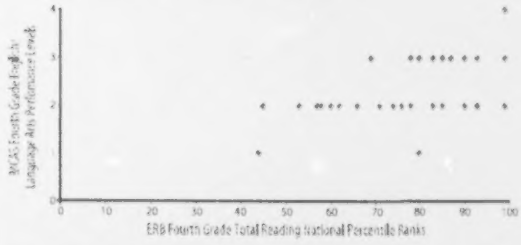
MCAS Eighth Grade Math Performance Levels Against Stanford 9 Eighth Grade Total Math National Percentile Ranks



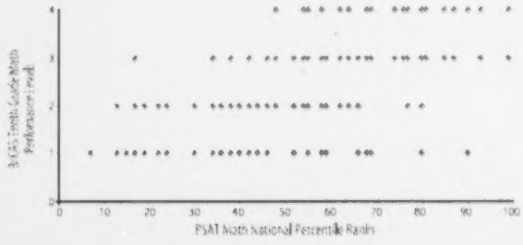
MCAS Eighth Grade Science/Technology Performance Levels Against Explore Eighth Grade Science National Percentile Ranks



MCAS Fourth Grade English/Language Arts Performance Levels Against ERB Fourth Grade Total Reading National Percentile Ranks



MCAS Tenth Grade Math Performance Levels Against PSAT Math National Percentile Ranks



BEST COPY AVAILABLE

the same range of scores on the second standardized test (found on the *x* axis) fall in all four MCAS categories. For example, from the 60th percentile (where students are 10 percentage points above the national average) to the 99th percentile on the Stanford 9, students fall into the four MCAS categories as follows: 270 students in Failure, 399 students in Needs Improvement, 411 students in Proficient, and 133 students in Advanced. Similarly, from the 60th to the 99th percentile on the 8th grade Explore science, students with similar percentile ratings are distributed throughout the four MCAS categories. Specifically, 11 students are classified Failure, 36 students Needs Improvement, 33 students Proficient and 2 students Advanced. For students at or above the 68th percentile on the 4th grade ERB reading in District C, 1 student is classified as a Failure (that person is, in fact, at the 80th percentile), 51 students are labeled Needs Improvement, 71 students are Proficient and 2 students Advanced on the 4th grade MCAS English/Language Arts test. Finally, the same picture we have seen three times previously appears in District D as well. When you look at just those students scoring at or above the 60th percentile on the PSAT math section, 6 students are in the lowest category, 6 students are in Needs Improvement, 64 students are in Proficient, and 65 are in the highest group on the MCAS test.

These overlapping scores across the four MCAS performance levels represent the variance that is not accounted for by performance on the second standardized tests. In statistical terms, variance accounted for is simply the squared correlation between two tests. For example, variation in District A students' scores on the Stanford 9 accounts for 70 percent (.84 squared) of the variation in their scores on the MCAS 8th grade math test. Variance unaccounted for, then, is the remainder of the total variation that is not accounted for in one test by another (30 percent, in this case). Other influences (variance unaccounted for which is discussed in the next section) affect the students' scores on one or both of the tests so that they do not match exactly.

Discussion

The MCAS scale scores, when compared with the percentile scores of external exams given to the same sets of students, seem to be consistent representations of performance in a given domain. The correlations, averaging around .7, indicate that the MCAS taps into similar but not identical information as do the other tests. Several questions are raised: What factors may be contributing to the correlations between the MCAS and other standardized tests and what might account for the unexplained variance in scores? Further, how are the correlations reflected in students' classifications into performance levels?

Many possible factors contribute to the less than perfect correlation between the MCAS and other standardized tests and to the overlaps in Figure 9. Some of these are listed below and discussed individually.

Possible Reasons for the Unexplained Variance in Scores on MCAS and Other Standardized Tests

- The MCAS measures similar but not identical domains to the other tests.
- Students did not take the MCAS or the second standardized exam seriously and thus did not score consistently.
- The time of year/length of the test may influence the students' test scores.
- The other standardized tests used are completely multiple-choice, while the MCAS includes open-response item types as well. The latter may have been unfamiliar to students and could have accounted for some of the unexplained variance.
- Unlike the other exams, the MCAS is a test of "world-class" standards and not an "off the shelf" standardized measure of achievement.
- Error on both tests may contribute to the unexplained variance in scores.

The MCAS measures similar but not identical domains to the other tests.

Part of the unexplained variance in the test scores may be due to the difference in content or skills related to the domains assessed by the various exams. The knowledge being tested on the MCAS English/Language Arts exam, for example, may not be exactly the same as that tested on the ERB reading test. Thus students' scores may differ, which would affect the correlation between the two tests. By way of illustration, the table below presents some of the content coverage of the MCAS English/Language Arts test and the Stanford 9 test.

Although similar in some respects the content being measured by the two tests is not identical. These differences may lead to varied performance on the two tests.

Stanford 9 Fourth Grade Reading Test	MCAS Fourth Grade English/Language Arts Test
• Students' understanding of directly stated details or relationships	• Students will acquire and use correctly an advanced reading vocabulary of English words, identifying meanings through an understanding of word relationships
• Students' comprehension of implicit information and relationships and their ability to make connections beyond the text	• Students will decode accurately and understand new words encountered in their reading materials, drawing on a variety of strategies as needed and then use the words accurately in writing
• Students' ability to analyze and evaluate explicit and implicit information and relationships	• Students will identify, analyze and apply knowledge of characteristics of different genres
• Students' ability to determine or describe strategies used by the writer or appropriate reader strategies to use in given situations	• Students will identify, analyze, and apply knowledge of theme in literature and provide evidence from the text to support their understanding

The Commonwealth's own test constructors (Advanced Systems in Measurement and Evaluation) use this type of information to suggest the validity of the MCAS. "The correlations found in [two studies]...indicat[e] that MCAS is appropriately related to other measures of achievement";[6] they were not intended to measure exactly the same information. According to the state, the correlations were "not too low, and not too high." This relationship, however, contributes to the overlap seen previously in Figure 9.

Students did not take the MCAS or the second standardized exam seriously and thus did not score consistently.

A second possible reason for the variance between students' scores on the MCAS and other standardized tests is the effort students put into taking the exams. MCAS scores in the first year of testing were not used for graduation decisions. For many tenth graders, up to three major standardized tests may have been required in a single academic year. But because no serious consequences were attached to their scores, students may not have been motivated to do well on these tests.

While there is some evidence that 8th and 10th grade students did not try to perform to the best of their ability on the MCAS, this cannot be said of the 4th grade. For example, the comparison of the 10th grade PSAT math percentiles with the MCAS math performance presented earlier reveals a substantial number of students with an MCAS scale score of 200 (the lowest possible score) across a range of PSAT scores. This "bottoming out" effect could be due to a lack of effort or motivation to do well on the MCAS. By contrast, 4th grade reading and math results across all districts show fewer students scoring at the lowest point. Some of the variance in test scores, then, may be attributable to lack of student motivation, but the notion does not seem generalizable across all grades and subjects tested.

Some of the variance in test scores, then, may be attributable to lack of student motivation, but the notion does not seem generalizable across all grades and subjects tested.

The time of year/length of the test may influence the correlation between students' test scores.

A third effect on the correlations and unexplained variance related to the motivation issue previously discussed may be the time of year and the length of the MCAS test administration. Unlike the other standardized tests, which typically take one to four days, MCAS testing takes the better part of two weeks. Students may have lost interest in trying their best after many days of tiring testing. Factors such as age and degree of attention can certainly affect performance; when students are asked to spend extended periods of time testing, the strain of the experience may well affect their scores. Also, the MCAS was administered in late April and early May, while the other tests came earlier in the school year (in the late fall or early spring). As discussed with respect to the lack of student effort due to no external motivation, students may also have been less inclined to concern themselves with testing so close to summer which would also explain some of the discrepancy in scores.

The other standardized tests used are completely multiple-choice, while the MCAS includes open-response item types as well.

Another possible influence on the correlations is the presence of open-response items on the MCAS. The other tests contained only multiple-choice items. Students' performance on the MCAS may differ from their performance on other standardized multiple-choice tests because they do not feel as comfortable with the open-item response type.

Another possible influence on the correlations is the presence of open-response items on the MCAS. The other tests contained only multiple-choice items. Students' performance on the MCAS may differ from their performance on other standardized multiple-choice tests because they do not feel as comfortable with the open-item response type.

Initial studies of this hypothesis, using one district's data, provide interesting findings. Using 4th grade MCAS English/Language Arts and Math scores, the following relationships emerged. First, the correlation between students' raw scores on the multiple-choice and on the open-response items on the MCAS was compared. The resulting statistics were .5 and .65 for the MCAS English/Language Arts and the Math tests respectively. Interestingly, while the English multiple-choice and open-response items were only moderately positively related, the math multiple choice and open-response scores had a somewhat stronger relationship. In practical terms, many students who did well on the multiple-choice section of the English/

Language Arts did not necessarily do well on the open-response and vice versa. The math correlation, however, suggests that, to a greater extent than with the English/Language Arts students who did well on one item type tended to well on the other.

To further examine the possibility that unfamiliarity with open-response items influenced correlations, we must also look at the relationships between MCAS multiple-choice and open-response and some external measure. Using the same 4th grade cohort discussed in the previous paragraph, the correlation between the MCAS open-response raw scores and the percentile ranks for the verbal and math ERB tests are .51 and .70 respectively (Figure 10).[7]

For the ERB verbal and math tests, the correlations between the MCAS multiple-choice scores and the ERB percentile ranks in these areas are .77 and .71 respectively (Figure 10). Again, the relationship between the MCAS English/Language Arts open-response scores and ERB verbal test is weaker than that between the MCAS Mathematics open-response items and the ERB math test. The multiple-choice correlations, however, are high for both tests. One possible reason for this discrepancy between English/Language Arts and math correlations may lie in the reading level of the 4th grade MCAS test. A study conducted by the Wellesley Elementary Literacy Department has shown that the reading level of the passages to which students

Figure 10

Correlations between MCAS Item Types and ERB Percentile Ranks

	ERB Verbal Percentile Rank	ERB Math Percentile Rank
MCAS English/Language Arts Open-response Raw Scores	.51	
MCAS English/Language Arts Multiple-choice Raw Scores	.77	
MCAS Mathematics Open-response Raw Scores		.70
MCAS Mathematics Multiple-choice Raw Scores		.71

responded was as high as ninth to tenth grade level. If the reading level was, in fact, well above grade level, MCAS English/Language Arts scores may not have accurately reflected student reading ability. The effect of reading levels beyond the students' capabilities may not have been present in the math section. Also, the open-response question type asked on the MCAS English/Language Arts test is markedly different from the type of open-response question on the Mathematics section. For example, Figure 11 presents an open-response item from both the Fourth Grade MCAS English/Language Arts and Mathematics tests given in the spring of 1998.

As can be seen from the examples given, the English/Language Arts open-response tended to call on students' strong understanding of a text to fully answer the questions, while

Figure 11

Fourth Grade MCAS English/Language Arts and Mathematics Open-response Items

Mathematics Open-response Question:

Nadine was playing a "Guess My Number" game with her mother. The first clue she told her mother was, "I am thinking of a three-digit whole number that has the digits 1, 5, and 8."

A. List all of the numbers that Nadine could be thinking of.

Nadine's next clue was this: "My number is also a multiple of 5."

B. List all of the numbers that Nadine could be thinking of now.

Finally Nadine told her mother the last clue: "When my number is rounded to the nearest hundred, it is 200."

C. What is Nadine's number?

D. Write three clues for another number game and number your clues. The game must have only one correct answer. Write the answer.

English/Language Arts Open-response Question:

What kind of person is Anastasia? Use information from the story in your answer. (From Chapter 1 of Anastasia by Lois Lowry.)

the mathematics open-response did not rely on any additional text beyond that in the question itself. Further investigation of this hypothesis is needed for more definitive results.

Unlike the other exams, the MCAS is a test of "world-class" standards and not an "off the shelf" standardized measure of achievement.

Yet another factor that might explain the discrepancy in students' scores is that the MCAS is harder than other commercially available standardized tests. (Alternatively, the cut scores are set unreasonably high.) In 1990, there was a groundswell for the establishment of "world-class" standards for the American educational system. The message of the National Education Goals Panel is illuminative:

> It is time to set our sights as high academically as we do athletically. We need to set world-class academic standards. They must be visible and reflected in curricula, instructional materials, teacher training, and assessment practices that enable our students to meet them and compete successfully with students of any country in the world. Not only should the top 5% of our students be as good as or better than the top 5% of students anywhere in the world, but the top 95% of our students should be as good as or better than the top 95% of students anywhere else. We must raise the expectations for every student in every school in the United States.

Source: http://www.negp.gov/issues/publication/negpdocs/6.html

Raising standards, the National Education Goals Panel pointed out, also makes the reform process difficult:

> Meeting these standards will not be easy. However, the standards are meant to define what students should aim for and, with sustained effort, be able to reach. It is a purpose that requires the commitment and effort of students, parents, teachers, administrators, government officials and members of the community. Schools need help. The purpose requires that we all accept responsibility for seeing that all our students reach a world-class level. We don't want to fool ourselves into thinking we have succeeded because our standards are set at low levels. As our National Education Goals state, we want students to succeed in challenging subject matter. Otherwise, America will remain a "nation at risk."

Source: http://www.negp.gov/issues/publication/negpdocs/6.html

Because the MCAS is ostensibly based on "world-class standards," the Massachusetts test is purported to be much more challenging than other "off the shelf" measures of achievement. We might expect students' scores to differ on the MCAS compared to other standardized tests because the tests are defining students' expected achievement not necessarily their present attainment.

Few would dispute the fact the MCAS is a challenging exam. The question, however, is whether Massachusetts students are well enough below "world-class" standards to support the differential performance on the tests discussed previously. There is some available evidence to suggest that this is not the case. Take, for example, the 1998 8th grade MCAS Science/Technology scores presented in Figure 12.

According to the results, over 70 percent of the 8th grade students in the Commonwealth of Massachusetts either failed or needed improvement in the area of Science/Technology, suggesting that students were not meeting the standards. National studies by the National Education Goals Panel (NEGP), however, raise questions about these low levels of achievement.

Figure 12

Percentages of Students in each of the Four Categories on the 1998 Eighth Grade MCAS Science/Technology Test

Performance Level	Percentage of Students Categorized in the Performance Level
Advanced	2
Proficient	26
Needs Impovement	31
Failing	40

*Percents do not add up to 100 as they are rounded to the nearest whole number

The NEGP uses the Third International Math and Science Study (TIMSS) results as a measure of the United States' progress toward Goal 5 (i.e., by the year 2000, United States students will be first in the world in mathematics and science achievement). According to their 1998 report, *Promising Practices: Progress Toward the Goals*, Massachusetts is one of only fourteen states that performed as well as or better in science than 40 of 41 nations (the exception being Singapore) on TIMSS as shown in Figure 13. This fact suggests that 8th grade students in the Commonwealth have a strong command of Science/Technology, strong enough to be "world-class." This finding, then, raises questions about how much of the unexplained variance in students' test scores on the MCAS and on other standardized tests can be attributed to the difficulty level of the MCAS and also about whether the performance level cut scores may be placed inappropriately.

Figure 13

Massachusetts Performance on the TIMSS Test

On the basis of a study linking state NAEP results to TIMSS in 8th grade science, 14 states would be expected to perform as well as or better than 40 out of 41 nations, including Canada, England, France, Germany, Hong Kong, Japan, Korea, and the Russian Federation. Only Singapore would be expected to outperform the following states:

Colorado	Connecticut	Iowa	Maine	Massachusetts
Minnesota	Montana	Nebraska	North Dakota	Oregon
Utah	Vermont	Wisconsin	Wyoming	

Also, although the MCAS may be, in fact, a "world-class" measure of students' academic abilities, the other tests districts have used to measure student achievement (as presented in this study) have a reputation for being academically challenging, as well. For example, the ERB is historically thought of as an extremely challenging exam and is most often used in suburban school districts and private schools to test students. The Stanford 9, too, has the reputation of being a well-developed, nationally respected assessment. The PSAT has been used for years as a predictive measure of student college success and as

a criterion in the awarding of the prestigious National Merit Scholarships. As such, each of these exams is externally considered quite challenging in its own right. The variance in student performance may be attributable, in part, to the difficulty of the MCAS. It is possible, however, that discrepancies in students' scores may also be due to unreasonably high cut scores.

Error on both tests may contribute to the unexplained variance in students' scores.

A final explanation for the resulting correlations and subsequent unexplained variance is error in the scores on the two tests. A student's score on a test is simply a single measure of his or her knowledge in a given domain. Given repeated testing, a student would not earn exactly the same score each time. Students may be tired, nervous, hungry, or bored on any administration, all factors that could potentially affect the results of the test. External factors such as room temperature, distractions in the hallways, lighting, and even eclipses[8] also impact test performance.

Regardless of how well it is made, an assessment (of achievement or some physical property) is never perfect.

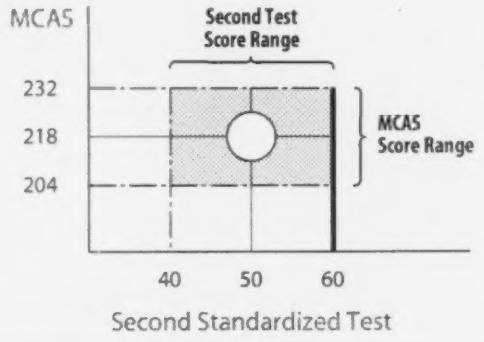
Most importantly, tests are fallible instruments. Regardless of how well it is made, an assessment (of achievement or some physical property) is never perfect. For example, the College of American Pathologists found that the error rate on cholesterol tests is, on average, 6.2 percent. This means that a person with a cholesterol level of 220 mg/dl "could expect a reading anywhere from 187, deep in the 'desirable' category, all the way to 267."[9] As another example, David Rogosa of Stanford University recently conducted a study using the Stanford 9 percentile ranks. According to his findings, a student who really belongs at the 50th percentile according to test norms will score within five points of that ranking on the test only 30 percent of the time in reading and 42 percent of the time in math. Said differently, 70 percent of the time, a student whose true score, as determined by statistical principles, is really at the 50th percentile on reading will be ranked more than five points above or below that point. That is a potentially substantial inaccuracy that could affect the results a student receives.[10] These examples serve as a warning that students' test scores on the MCAS and/or the other standardized test may be filled with error that does not allow for exact comparisons.

It is safe to assume that the MCAS and the commercially developed standardized tests have been created with care and that the obtained scores are reasonable estimations of a student's achievement. However, all tests are fallible estimators of achievement. As Rogosa's study indicates even well constructed tests can provide misleading information to parents and students. Although all of these tests seem to contain minimal error, even small amounts of error can greatly impact individual students.

For example, Figure 14 presents a student's scores on the MCAS and a second standardized test (represented by the circle). Assume the error on each of the tests is represented by the dashed lines. The error on the 8th Grade MCAS math test is such that a student's actual score might be ± 14.4 points from the obtained score when a student has a scale score below 240. A student with a score of 218 on the MCAS English/Language Arts exam, then, might have a "real" score as low as 204 and as high as 232. Further, assume that the standard error on the second standardized test is ± 5 points. With 95 percent confidence, a person at the 50th percentile, then, could in fact be at the 40th to 60th percentile. The shaded area, then, represents the full range of potential real scores that this student could achieve. It is important to note how error on both tests can affect the ways in which test scores are interpreted. As the example above showed, a student could move from Failing to Needs Improvement on the MCAS as well as 10 percentile points on the second test. This example serves to highlight the realistic effects that error may have on test scores; small differences in the "real" test scores of students may carry a large impact.

Figure 14

Hypothetical Student Responses and Error Bands for MCAS Test and a Second Standardized Test



Correlations and Performance Levels

In considering some of the factors influencing the relationship between the MCAS and other standardized tests, then, the resulting correlations across subjects, grades, and districts seem to be reasonable. The MCAS was not designed to be strictly parallel to any of the other tests discussed; a perfect correlation between the two measures would indicate that one is, in fact, not necessary. Instead, correlations suggest both tests provide additional or corollary information that may be useful to students, parents, and teachers.

Given these existing correlations, some students and parents will receive mixed messages related to their academic achievement on the MCAS versus some other standardized test(as evidenced in Figure 9). Specifically, some students' MCAS performance level results and their scores on other standardized tests will be discrepant. It is important to remember that the location of the four performance level cut points on the MCAS forces breaks in the continuum of scale scores that may not ideally categorize students; moving the cut points might give distinctly different results.

It is important to remember that the location of the four performance level cut points on the MCAS forces breaks in the continuum of scale scores that may not ideally categorize students; moving the cut points might give distinctly different results.

What lesson should people draw from this, then? Tests are not infallible and cut scores increase the fallibility of interpretation. The MCAS performance levels are not an unquestionable source of information about students' performance in the domains of English/Language Arts, Math, and Science/Technology. Cut scores are risky; different cut points might provide drastically different information to students about their accomplishments. Also, other standardized tests (as well as other information, e.g. teacher-made tests, grades) give supplementary and valid information related to these areas. No one measure tells us everything.

Many people in positions of power (e.g. politicians) believe it is important to use the MCAS as a measure of Massachusetts students' achievement in the areas of English/Language Arts, Mathematics, and Science/Technology as the driving element in the reform effort. Recently, Governor Paul Cellucci made a public pronouncement that he would publicly rank districts in the Commonwealth based on MCAS performance. The grading system, as Massachusetts continues toward a comprehensive

accountability system, would apparently reflect the combination of schools' overall performance on the MCAS scores and their level of annual improvement. We would hope that the issue of error would also be considered as annual improvement is determined. However, how this should be done is not clear.

Regardless of problems that exist (e.g., error), the MCAS tests are effective and powerful tools in Massachusetts educational reform. Related to that reform effort, two points are important to remember. First, some form of external evidence of the legitimacy of the MCAS scores is needed to ensure that performance levels are fair and accurate measures of student performance. Second, making important decisions based on a single measure is unadvisable. The Standards for Educational and Psychological Testing (1999) state that "in educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision".[11] No single test should determine whether a student graduates from high school or moves to another grade.

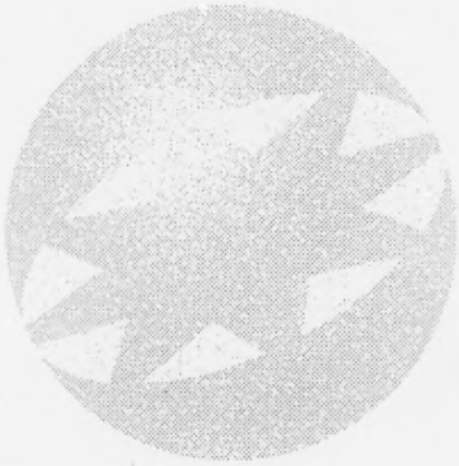
Conclusion

Performance levels are enticing. Parents and teachers can look at a score report and see whether Amy is "proficient" in science or "needs improvement" in math. Using these labels as criteria, schools or districts may be judged by how many students achieve "proficient" or "excellent" status on a test. This profound simplicity is what makes performance levels such desirable tools. Unfortunately, this is also what makes them so worrisome, especially when high stakes decisions are involved.

The public places great faith in the infallibility of test results. By extension, people tend to view performance levels the same way. In this case, simplicity diverts attention away from a fundamental problem. Performance levels are based on cut scores. Cut scores, in turn, are based on judgment. The point is, no matter what procedure or combination of procedures is used, *judgment* is a key element in all of them. Of course, if judges' decisions are accepted as final and unquestionable, the story ends. The problem is, as long as there is judgment involved in

the cut-score setting procedure, we can never be completely sure performance levels accurately reflect student achievement. At the very least, we have to use other kinds of evidence to try and evaluate whether choices based on performance levels are appropriate. Sometimes, as we saw with the MCAS data, students may score well above average on a commercially developed standardized test or on a "world-class" test like TIMSS and still be labeled a "failure" or in "need of improvement". These sorts of mixed messages are all the more troubling in light of the current use of performance levels to make high stakes decisions.

Most of us have experienced the stress associated with taking a test. But what if your future is directly tied to whether or not you pass? Add the fact that the definition of "pass" is largely a judgment call. It is clear that mislabeling a student in this situation has emotional and psychological consequences beyond the prospect of graduation or promotion decisions. However, there is no indication this use of performance levels will change. In fact, an increasing number of students face the prospect of having to pass one of these high stakes tests in the near future. Using the MCAS results, we have already seen the impact of performance levels on perceived student achievement. Now imagine that a mislabeled student is barred from graduating. How would you explain this to a parent? Given these circumstances, it is imperative that we continue to scrutinize performance levels to help insure their proper use.



notes

1 Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), pp. 237-261.

2 Kane, M. (1994). Validating the performance standards associated with cut scores. *Review of Educational Research*, 64(3), pp. 425-461.

3 A percentile rank is a student's position in a group relative to the percentage of group members scoring at or below that student's raw score. For example, a student at the 90th percentile scored higher than 90 percent of the test takers in the norm group (the group of students who were initially tested to produce the scores that describe the test performance of a national sample of students).

4 A caveat is necessary here. The correlations presented here and throughout the paper are specific to the districts, students, and tests discussed; if all the districts had used the same commercially developed standardized tests to assess their students, for example, correlations with the MCAS would be slightly different for each. Because four different standardized tests were used, we must be careful in our generalizations.

5 Most standardized tests correlate at around .7 to .8 between similar domains (see 1975 Anchor Test Study by the National Center for Education Statistics). As an example, the Stanford 9 uses correlations with the Otis-Lennon School Ability Test to provide evidence of validity of the Stanford 9. Their obtained correlations range from .64 to .77 across various grades.

6 Advanced Systems in Measurement and Evaluation (1999). *Massachusetts Comprehensive Assessment System 1998 MCAS Technical Report*. MA: Advanced Systems.

7 There were 28 multiple-choice and 6 short-answer and extended-response questions on the 4th grade MCAS English/Language Arts section. The MCAS math section included 21 multiple-choice and 11 short-answer and open-response questions. The weighting of the short-answer and open-response questions on both tests was such that those sections were worth more total points than the multiple-choice questions. Correlations range from .64 to .77 across various grades.

8 In El Paso, Texas, district administrators requested a change of test date for the TAAS due to a solar eclipse occurring on one of the test dates. They were worried that student performance might be impeded.

9 Moore, T.J. (1989). The cholesterol myth. *The Atlantic Monthly*, 126(3), pp. 37-60.

10 For a more detailed discussion of this study, see CRESST web page: http://www.cse.ucla.edu/CRESST/Reports/drrguide.html

11 American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

NBETPP Home
Go To:
http://nbetpp.bc.edu
NBETPP
Visit us on the World Wide Web at nbetpp.bc.edu for more articles, the latest educational testing news, and information on NBETPP.

BEST COPY AVAILABLE

About the National Board on Educational Testing and Public Policy

Created as an independent monitoring system for assessment in America, the National Board on Educational Testing and Public Policy is located in the Peter S. and Carolyn A. Lynch School of Education at Boston College. The National Board provides research-based test information for policy decision making, with special attention to groups historically underserved by the educational systems of our country. Specifically, the National Board

- Monitors testing programs, policies, and products
- Evaluates the benefits and costs of testing programs in operation
- Assesses the extent to which professional standards for test development and use are met in practice

This National Board publication series is supported by a grant from the Ford Foundation.

The National Board on Educational Testing and Public Policy
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Telephone: (617)552-4521 • Fax: (617)552-8419
Email: nbetpp@bc.edu

Visit our website at **nbetpp.bc.edu** for more articles, the latest educational news, and for more information about NBETPP.

NBETPP

The Board of Directors

Peter Lynch
Vice Chairman
Fidelity Management and Research

Paul LeMahieu
Superintendent of Education
State of Hawaii

Donald Stewart
President and CEO
The Chicago Community Trust

Antonia Hernandez
President and General Council
Mexican American Legal Defense and Educational Fund

Faith Smith
President
Native American Educational Services

BOSTON COLLEGE